From Categories to Individuals in Real Time — A Unified Boosting Approach

David Hall Pietro Perona
California Institute of Technology
dhall,perona@vision.caltech.edu

Abstract

A method for online, real-time learning of individual-object detectors is presented. Starting with a pre-trained boosted category detector, an individual-object detector is trained with near-zero computational cost. The individual detector is obtained by using the same feature cascade as the category detector along with elementary manipulations of the thresholds of the weak classifiers. This is ideal for online operation on a video stream or for interactive learning. Applications addressed by this technique are reidentification and individual tracking. Experiments on four challenging pedestrian and face datasets indicate that it is indeed possible to learn identity classifiers in real-time; besides being faster-trained, our classifier has better detection rates than previous methods on two of the datasets.

1. Introduction

Detecting objects in image collections and video is a rich area of application of visual recognition. Technical approaches change significantly depending on whether one focuses on *individual-objects* [35] or on *categories* [6, 16, 19] and the two challenges are pursued as distinct research questions. While this separation is useful in academic research, real-world systems require a combination of category and individual detection.

For concreteness, we describe two such scenarios. The first is *tracking*. Applications include tracking of pedestrians in railway stations and airports, vehicles on the road for traffic monitoring and faces for interfacing people with computers. A common approach used is tracking-by-repeated-detection. In its simplest form this consists of a frame-by-frame category detector followed by an algorithm that combines detections across space and time into trajectories [3, 5]. Trajectory smoothness constraints confer a degree of robustness to false detections; the main challenge is continuing trajectories when detection fails over multiple frames because of occlusion, unusual pose or unfavourable lighting conditions. Once an object (a pedestrian) has been detected by the category detector (or by a human operator), tracking is made more robust by training an individual-object detector exploiting the specifics of the individual's appearance (a person with a red sweater and backpack) [38].

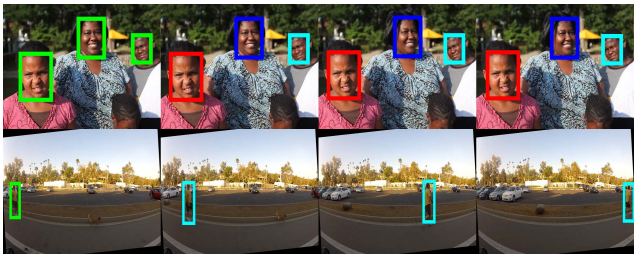

Figure 1. Tracking individuals across a video sequence: faces (top) and pedestrians (bottom). The first column shows the detections made by a category detector. These detections are used to initialise three individual face detectors, and a single individual pedestrian detector, and are then evaluated on the subsequent frames in the video sequence. Individual detector outputs are colour-coded.

Reidentification is another scenario with applications in video surveillance [10] and content-based indexing of image collections, consumer videos and commercial video libraries. After a category detector detects instances of the category, individual detectors are trained to cluster and classify the individuals that appear in the collection [17]. Individual reidentification across networks of cameras is similarly important [7, 8, 28, 46].

It is clear from these examples that it is useful to detect objects both as members of categories and as individuals. In the first scenario, individual detectors trained on-the-fly improve tracking robustness. In the second scenario, individual classifiers reveal recurring individuals in an entire collection or video stream. In both cases it is crucial that an individual detector is trained in real-time and that its runtime cost does not add significantly to the overall computational cost of the system.

Here we present a method for real-time, online training of individual detectors from individuals that are detected by a category detector. We make three main contributions:
1. A unified boosting-based approach for simultaneous category and individual detection.
2. A method for training individual detectors in real-time from a single training example.
3. Two novel challenging datasets of faces and pedestrians.

2. Related work

Researchers in visual categorisation agree that objects are best represented as constellations of visually distinctive parts that appear in flexible geometrical arrangements [21, 33, 6, 19]. A variety of practical approaches to detecting parts and representing mutual positions have been proposed, where the representation of shape is either explicit [20, 19] or implicit [34, 39, 43, 14]; best performance is currently obtained with discriminatively trained part detectors [14, 19]. This work is based on boosted cascades of classifiers [43, 4, 14] because they deliver state-of-the-art detection performance at video-rate computational speeds [2].

Researchers focusing on individual detection [35] and re-identification [10] focus both on the design of (domain-specific) features [28, 7, 46] and on efficient algorithms for detection and classification [35]. In our work we are feature-agnostic, in that our framework allows the implementation of a large variety of different features, and we rely on the computational efficiency of cascaded boosted classifiers.

Online learning of detectors for tracking individual objects, given an operator-supplied initial training window, is a topic of much interest [9, 31]; the main challenge is *drifting* from the original target. The closest work to our own are the online boosted trackers of Grabner *et al*. [24, 23, 25]. In their work, boosted individual-object detectors are trained online and are paired with a *prior* to limit drift. The individual detectors operate at frame rates of between 10–15 frames-per-second on a video with a resolution of 640x480; however, this cost is in addition to the cost of running a category detector, the output of which initialises the individual detector.

Our work aims to produce a unified approach for simultaneous category and individual detection to ensure that real-time operation can be achieved. We focus on two questions that have, to our knowledge, not yet been studied: assuming that a category detector is available, (a) how to design individual detectors whose *additional* run-time cost is small or zero; (b) how to train such individual detectors on-the-fly with *minimal computational cost* once one or more training examples become available from the category detector.

3. Approach

Our approach is based on using *cascaded boosted classifiers* both for category and individual detection [42, 43, 14, 13]. Detectors of this form have been shown to be fast and have state-of-the-art detection performance [2]. In order to make this paper self-contained, we review cascaded boosted classifiers (Sec. 3.1); discuss the implementation of category detectors (Sec. 3.2) and finally outline the approach for designing individual detectors (Sec. 3.3).

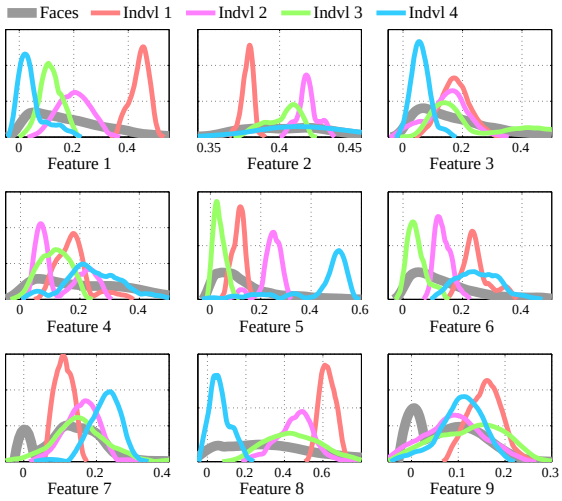

Figure 2. The empirical distributions of a set of faces and four individual faces across the first nine features selected by AdaBoost for the category detector. The set of faces was used to train the category detector (refer to Sec. 4 for details). The four individuals represent a single sequence of an individual lasting for at least fifty-five frames from a test video in the FPOQ dataset (Sec. 4).

3.1. Boosting

A boosted classifier takes feature vector $\mathbf{x} \in \mathbb{R}^D$ as input and outputs a binary decision:

$$H(\mathbf{x}) = \text{sign}\left(\sum_{m=1}^{M} \alpha_m h_m(\mathbf{x}) - \tau\right) \tag{1}$$

where the threshold τ is chosen to produce the desired tradeoff between false reject rate and false alarm rate. Given a labelled training set $\{\mathbf{x}_i, y_i\}_i$ of N samples, the boosted classifier is trained by greedily minimising a loss function (which depends on the type of boosting being used: AdaBoost, LogitBoost, *etc*.). This means that at each iteration m up until the maximum number of iterations M an optimal weak classifier $h_m(\mathbf{x})$ and weight α_m are selected. For training, each data sample x_i is assigned a weight w_i^m (depends on the loss function). At each iteration, samples that are classified incorrectly are weighted more heavily which means the penalty for classifying them incorrectly in subsequent iterations increases.

3.2. Category Detector

In this work category detectors are trained offline using AdaBoost [22]. The family of weak classifiers used are stumps. This means that given an input $\mathbf{x} \in \mathbb{R}^D$, the decision only depends on the j-th dimension of $\mathbf{x}$, a threshold $\theta \in \mathbb{R}$ and a polarity $p \in \{\pm 1\}$

$$h_m(\mathbf{x}) = \begin{cases} 1, & p_m x_{j_m} > p_m \theta_m \\ -1, & \text{otherwise} \end{cases} \tag{2}$$

During training, the optimal weak classifier at the m-th iteration of boosting is selected by choosing j, θ and p so that the number of the N weighted training examples that are misclassified is minimised. Choosing these parameters at each of the M iterations is $\mathcal{O}(MND)$ and is the most computationally expensive part of training a boosted classifier. Note that any boosting method and decision trees of any depth could be used to train the category detector; our proposed method is agnostic to these choices. For the sake of clarity, in the following discussion we will continue to refer to AdaBoost and decision stumps.

3.3. Individual Detectors

The proposed approach for designing individual detectors relies on four key principles: 1) the individual detector has the form of a cascade of boosted classifiers (Eqn. 1); 2) an individual detector is learnt from a single instance of the individual; 3) the *training* and 4) the *runtime* costs of an individual detector must be minimal to guarantee online, real-time operation.

The most obvious strategy for training an individual detector is to repeat the AdaBoost training process using an object identified by the category detector as a positive training example. Recent work has made the training stage of AdaBoost faster [1]; however, it is still a computationally expensive process and remains ill-suited for real-time operation. In the following exposition we will look at the limitations of a traditional boosting approach and examine a set of constraints that can be placed on the individual detectors to avoid the computationally costly steps of a traditional boosted detector.

The first design goal requires individual detectors to be of the same form as the category detector. This is a reasonable restriction to place on the individual detectors since cascades are fast, making them suitable for real-time operation and their performance is state-of-the-art as has been previously mentioned. This requirement also ensures there is simplicity in design and a unified approach for both category and individual detection.

The second principle, that an individual detector is learnt from a single instance of an individual is also a reasonable requirement. Training using a traditional boosting approach is possible by jittering or transforming the original example. This results in multiple, slightly altered versions of the original instance which can all be used as positive training examples. The drawback here is that negative examples are now required. This either requires precomputed negatives to be stored in memory (which may be limited) or for negative examples to be mined online which is another costly computation.

To ensure design goals 3) and 4) are satisfied it is important to examine the most computationally expensive steps in the object detection pipeline. The first of which is feature computation.

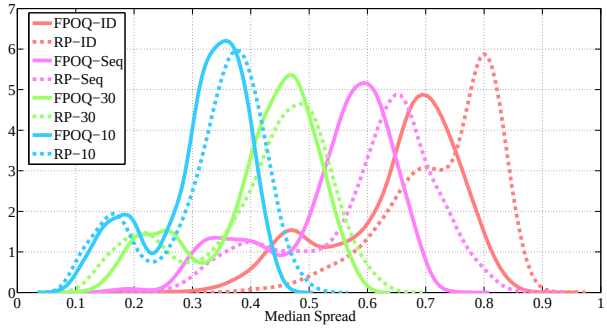

Figure 3. The empirical distributions of the median spread (Eqn. 3) for the M features selected by a boosted category detector of the faces in the FPOQ (solid) and of the pedestrians in the CRP (dashed) datasets (Sec. 4). The red curves correspond to the median spread of individuals across multiple sequences; the pink, to individuals in a single sequence; the green, to 30 consecutive frames of an individual and the blue to 10 consecutive frames of an individual.

Computing features is expensive; however, some features have already been computed for the category detector. If we can re-use the same features for the individual detector, the additional runtime and training costs for an individual detector due to feature computation are zero. For this reason, **individual detectors will be constrained to only use features that have already been computed for category detection**.

A second computationally expensive stage to consider is feature selection during training. This is equivalent to choosing parameter j at the m-th boosting iteration (Sec. 3.2). Typically, D, the dimensionality of the feature space is large so performing this optimisation is costly; however, this optimisation can be avoided if the **individual detectors are constrained to use only the M features that were selected by AdaBoost for the category detector**. We will denote this set of features by $\mathbf{J} = (j_1, \ldots, j_M)$ where $j_m \in \{1, \ldots, D\}$ and the importance of each feature through the weights $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$. The additional training cost due to feature selection is thus zero.

It is not intuitive that category detection features (features that are good at distinguishing faces from background), are also useful for individual identification (distinguishing faces from other faces). It is reasonable to expect that the features AdaBoost would select for a face detector are features that are *common* to all faces; consequently, these common features should be uninformative for distinguishing *between* faces. Consequently, constraining the individual detectors to only use the M features of the category detector ought to doom it to failure. However, this intuition is not necessarily correct.

For a category detector to perform well, it needs to be able to detect many different types of faces in different lighting conditions. It is not necessary for an individual stump

to cover the complete range of feature values; it only needs to capture a narrow range. Breadth is achieved by combining multiple stumps. It would then be reasonable to expect that the feature distributions for an individual are localised within narrower intervals that are contained within the category distribution for that particular feature.

Plots in Figure 2 show the empirical distribution of nine features from faces used to train the category detector (grey) as well as the distributions for four different individuals (colour); the individual distributions are obtained from video sequences lasting at least fifty-five frames from a test video in the FPOQ dataset (Sec. 4). (In video, a sequence is a consecutive set of frames in which an individual appears. Individuals occur in multiple sequences across the length of the video.) Each subplot is for one of the first nine features selected by AdaBoost for the category detector. This plot suggests that the distribution of a feature for a particular individual is localised within the broader category distribution. To substantiate this claim, a statistical analysis across many different individuals is required.

Let $X^+ \in \mathbb{R}^{N\text{x}M}$ be the matrix of M-dimensional feature vectors of the N positive examples of a category. The N positive examples consist of multiple instances of the same individual, possibly under different pose and lighting conditions. The range of feature j for the *entire category* can then be defined as $r_j^+ = Q_{0.95}(X_j) - Q_{0.05}(X_j)$ where Q_p is the p-th quantile and X_j is the j-th column of X^+. The range of feature j for an *individual across multiple sequences* is $r_j^i = Q_{0.95}(X_j^i) - Q_{0.05}(X_j^i)$ where X_j^i is the j-th column of X^+ but only with the rows that correspond to individual i. The median spread l_j across all individuals for feature j is then defined as:

$$l_j = \underset{i}{\text{median}}\left(\frac{r_j^i}{r_j^+}\right) \tag{3}$$

The median spread of an individual in a single sequence, in 30 consecutive frames and in 10 consecutive frames is also considered and can be defined similarly. Figure 3 shows the empirical distributions of the median spread for the features for all faces in the FPOQ and all pedestrians in the CRP datasets (Sec. 4) .

Figure 3 suggests that reidentifying individuals within sequences (pink curves) or within 30 (green curves) or 10 (blue curves) frames of each other is possible since most features are localised (the spread is small with respect to the category distribution). It also suggests that reidentifying individuals across sequences (red curves) may be problematic since there are more features with higher spread values; however, the architecture of a cascaded boosted classifier provides some robustness to this variability between sequences. If there are enough features that exhibit limited variability (FPOQ (solid red) has a number of features with a spread of less than 0.5) then an individual detector

Figure 4. (Left) The faces of five different people from the FPOQ dataset. (Right) Examples of five different pedestrians (one for each row) from the Roadside Pedestrian Dataset. The individuals were sampled randomly from a video in each of the datasets. Faces are quasi-frontal; lighting varies between overcast and direct sunlight; a variety of expressions are present as individuals are filmed while talking. Pedestrians show a wide range of poses, lighting conditions and backgrounds.

may still classify an individual correctly across sequences because there is sufficient evidence to suggest that the individual is present.

The third and final computationally expensive stage in a traditional boosting approach is threshold selection during training. Even if features have already been selected (parameter j has been fixed), the optimal threshold θ, at the m-th boosting iteration still needs to be chosen (Sec. 3.2). This is once again computationally expensive; however, this optimisation can be avoided if we consider an alternative approach.

Selecting the thresholds for a single weak classifier $h'(x')$, which depends on a single feature $x' \in \mathbb{R}$, can be achieved at almost zero computational cost by using transfer learning. Consider the single instance of an individual that has been detected by the category detector and call γ' the value of feature x' for this instance. Figure 2 suggests that the distribution of features for an individual tends to be localised. The average spread σ' of feature x' across many individuals may be estimated offline using a validation set composed of images grouped by individual. An interval $(\gamma' - \beta\sigma', \gamma' + \beta\sigma')$ can then be defined where β is a free parameter that can be tuned experimentally. This interval represents the most likely values that the feature x' will take for the individual detected by the category detector. According to this strategy, the weak classifier $h'(x')$ may be obtained directly from one training example:

$$h'(x';\gamma',\sigma') = \begin{cases} 1 & \gamma' - \beta\sigma' < x' < \gamma' + \beta\sigma' \\ -1 & \text{otherwise.} \end{cases} \tag{4}$$

This weak classifier provides evidence for an individual being present (absent) if the feature x' lies inside (outside) the

interval $(\gamma' - \beta\sigma', \gamma' + \beta\sigma')$. The training cost for selecting the thresholds for a single weak classifier is thus a small constant.

Using the ideas presented, an individual detector in the form of a cascaded boosted classifier can now be constructed. Given the set of features $\mathbf{J}$ and weights $\boldsymbol{\alpha}$ that were selected for the category detector, the first instance $\mathbf{u}^k$ of the k-th individual detected by the category detector, and an estimate of the spread $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_M)$ of the features $\mathbf{J}$ a classifier $F^k(\mathbf{x})$ for the k-th individual can be defined by:

$$F^k(\mathbf{x}; \boldsymbol{\gamma}^k, \boldsymbol{\sigma}) = \sum_{m=1}^{M} \alpha_m h'(x_{j_m}; \gamma_m^k, \sigma_m) \qquad (5)$$

where $\boldsymbol{\gamma}^k = (\gamma_1^k \ldots \gamma_M^k)$ with $\gamma_m^k = u_{j_m}^k$. The total computational cost for learning an individual detector using the outlined approach is only $\mathcal{O}(M)$. This is significantly less expensive than if the individual detector was trained using standard AdaBoost which is $\mathcal{O}(DMN)$.

4. Experiments

To assess the performance of our procedure for training individual detectors (we call it IDBoost), as well as to determine the limitations and applicability of the approach, two types of experiments, that illustrate two possible operating regimes for the individual detectors, are considered. The experiments are evaluated on two different categories: faces and pedestrians. All experiments are carried out using the multi-scale detection framework of Dollar [13], with the channels of brightness, colour, gradient magnitude and gradient orientation used as features; the code is available in Dollar's publicly available Image and Video Matlab Toolbox. In all experiments, the parameter $\beta = 0.6$; this choice ensured detectors operated at a fast enough rate whilst maintaining performance.

4.1. Datasets

To carry out the experiments it was necessary to collect two new challenging video datasets that contain many different individuals that reappear at different moments in time.

The first dataset is the Fifty People One Question (FPOQ) face dataset. It contains 6 videos with 222 annotated individuals across 725 sequences (in video, a sequence is a consecutive set of frames in which an individual appears). Each annotation contains the bounding box, the identity and the sequence number of the face. In total the are 68,676 bounding boxes; 78,181 frames; and 57,274 frames that contain faces. The videos were collected from YouTube (*e.g.* http://youtu.be/csHddXn91YE) and involve either a single individual or groups of individuals being asked a question in front of a fixed camera. Their responses are edited in such a way so that an individual's response is interspersed between the responses of others. This means individuals can appear at any time point within the video. Examples of the different individuals as well as the different appearances of those individuals throughout the video are displayed in Figure 4. The face category detector was trained using 800 different faces extracted from single frames across 26 other videos (these videos are in the same style as the FPOQ videos). The faces used to train the face detector are not used during testing. The spread $\boldsymbol{\sigma}$ of the features selected by the face detector averaged over many individuals is estimated from a video in the FPOQ dataset. The other 5 videos are used for testing.

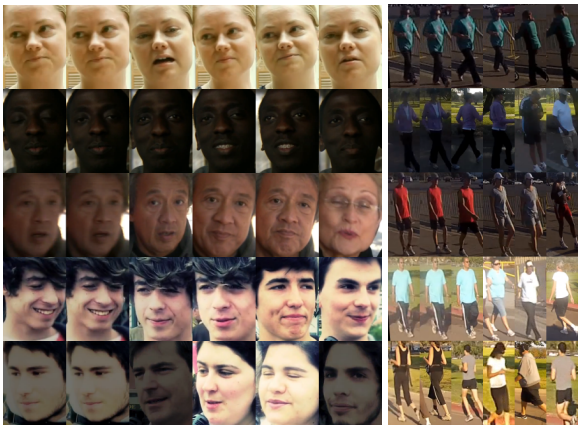

Figure 5. A demonstration of the reidentification of individuals in (left) the FPOQ and (right) the CRP datasets using IDBoost. Column 1 shows the instance used to train the individual detector, it is the first instance of that individual detected by the category detector. Columns 2–6 show the top, 50th, 100th, 200th and 300th best scoring results returned by the individual detector after being evaluated on the ∼10,000 detections made by the category detector. For the CRP dataset columns 2–6 show the top, 5th, 10th, 20th and 30th best scoring results from ∼2500 detections.

The second dataset is the Caltech Roadside Pedestrian (CRP) dataset. It contains 2 videos with 170 annotated individuals across 263 sequences. In total there are 7450 bounding boxes; 77,450 frames; and 5606 frames that contain pedestrians. Each video is captured by mounting a rightwards-pointing video camera to the roof of a car. The car then completes two laps of a ring road within a park where there are many walkers and joggers. This dataset is more challenging than the face dataset due to the considerable differences in lighting and pose for an individual. Figure 4 displays a few examples of the pedestrians in this set. The pedestrian category detector was trained using 64 of the individuals in the CRP dataset; $\boldsymbol{\sigma}$ was also estimated from this set. the remaining 106 individuals were used for testing.

For completeness we also evaluate our method on two existing re-identification datasets, VIPeR [26], which contains 632 person image pairs and ETHZ [41, 15] which con-

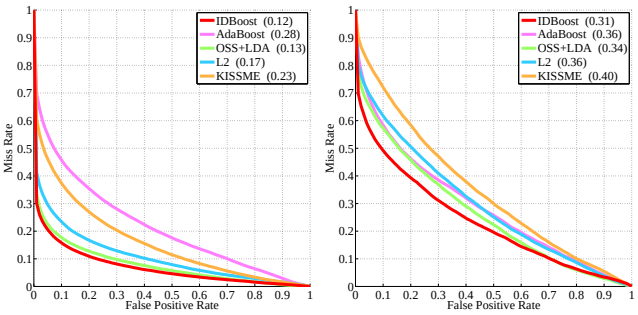

Figure 6. Reidentification performance. ROC curves for reidentification averaged over the (left) 207 individuals in the FPOQ dataset and the (right) 106 individuals in the CRP dataset. The category detector extracts all instances of an object in the video (in the FPOQ dataset there are 9375–15521 face detections per video; in the CRP dataset there are ∼2500 pedestrians detected per video). An individual detector is then trained (using either IDBoost, AdaBoost, OSS+LDA, L2 or KISSME) for each of the individuals present in the video using the very first instance of that individual. Each detector is then applied to every other instance that was detected by the category detector in the video. An roc curve is generated for each of the individual detectors and the average roc across all individuals is then computed. A true positive occurs when an individual detector fires on the same individual that it was trained on. The mean equal error rate is also given for each of the methods. Our method has the best performance on both datasets.

tains multiple windows of people extracted from 3 video sequences. For conciseness we only report the results on ETHZ Sequence 2 which contains 35 persons across 1961 images.

4.2. Reidentification

Individual detectors can operate in two modes, the first is as a classifier, evaluated on the single windows indicated by the category detector to contain an object of interest. The reidentification problem involves reidentifying instances of an individual (each instance varies in pose, lighting, background and occlusion) from a set of many different individuals. The reidentification experiments are run on both the FPOQ and the CRP datasets. Each new individual is determined by a human operator, an individual detector is learnt using this example and it is then evaluated on all detections made by the category detector (even the category detections that occur in the frames prior to the individual detector being created). Experiments are also run on VIPeR and ETHZ SEQ2 to examine the cross-dataset performance of IDBoost (CRP is used for training, VIPeR/ETHZ for testing).

We compare the performance of individual detectors trained using our method to individual detectors trained using three other methods: AdaBoost [22]; the One-Shot Similarity score using LDA (OSS+LDA) [45] (code obtained from the authors website); and an l^2 distance as a baseline. These methods use a single example for training to

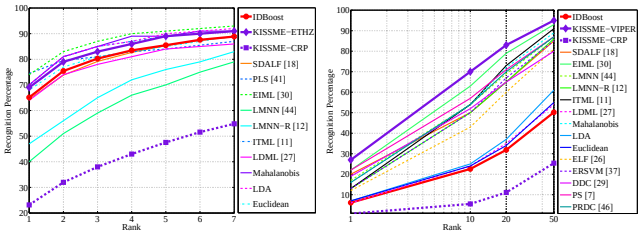

Figure 7. Cross-dataset performance of our method (IDBoost) (red curves). IDBoost was trained (learning σ (Eqn 4)) using the validation subset of CRP and tested on ETHZ SEQ2 (left) and VIPeR (right). For reference we provide the performance curves of a number of methods that were trained on those datasets[18, 40]. ID boost performs comparably to the other methods on the ETHZ dataset, whose statistics are close to CRP's since individuals were sampled by a moving camera. On the VIPeR dataset IDBoost performs less well than methods which were trained on VIPeR; this is probably because individuals were imaged by separate cameras with different lighting conditions. For a fairer comparison, we take the state-of-the-art method KISSME [32] and train it using the validation subset of CRP. KISSME-CRP performs poorly on both VIPeR and ETHZ. It over-fits the training data, so is unable to generalise to new datasets with different statistics. This result suggests that IDBoost has a greater capacity to generalise across datasets.

provide a fair comparison. We also make a comparison to KISSME [32] a metric learning algorithm that has state-of-the-art performance on the VIPeR dataset (code obtained from the authors website). To learn an individual detector from a single example using AdaBoost, virtual positives are created by applying slight transformations to the example whilst negative examples are sampled from the background of the frame; this is similar to the initialisation stage of the online boosting trackers [24, 25]. KISSME is trained using the same FPOQ and CRP validation subsets that IDBoost is trained on unless otherwise specified.

The results of our experiments are in Figure 6 with examples in Figure 5. They indicate that reidentification of faces is fairly easy due to the interview style of the videos with the pose, background and lighting of the face changing minimally so an individual looks the same from the first instance to its last. Reidentifying pedestrians is much more difficult due to the large changes in appearance that occur due to lighting and pose. From the ROC curves of Figure 6 it is clear that our method (IDBoost) performs equally or better than any of the other methods, despite the fact that its computational cost is a tiny fraction as shown in Figure 9 (top). Figure 7 gives the results of cross-dataset performance with IDBoost doing significantly better than KISSME-CRP.

4.3. Tracking

An individual detector can also operate as a sliding-window detector, evaluated on every window in every frame of a video. The additional runtime cost in this mode of operation is higher than in the reidentification scenario since

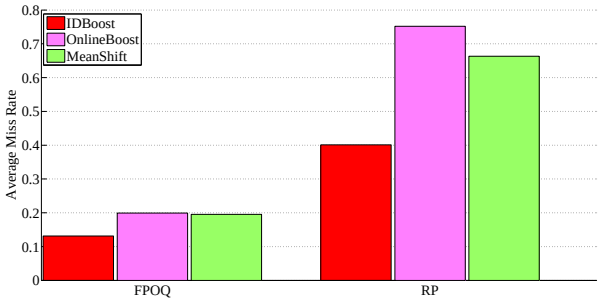

Figure 8. Tracking performance. The miss rate averaged over the (left) 680 sequences of individuals in the FPOQ dataset and over the (right) 199 sequences of individuals in the CRP dataset. For each sequence, the output of the category detector evaluated on the first frame of the sequence is used to initialise an individual tracker (either an IDBoost, OnlineBoost or MeanShift tracker). The miss rate (the number of times the tracker misses the individual it was initialised to track) is then computed for each sequence and the average miss rate over all sequences is the computed. Our method has the best performance on the both datasets.

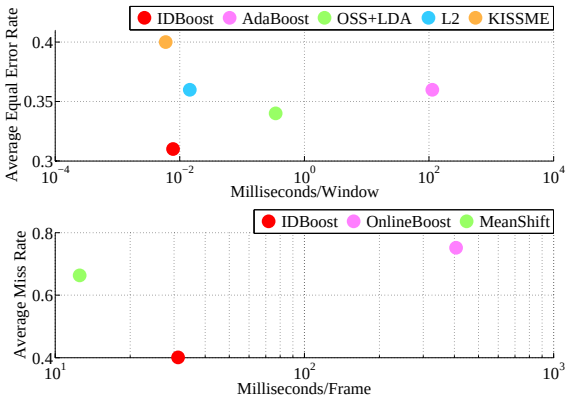

Figure 9. Computational performance. The average time it takes to train and evaluate an individual detector (either per window or per frame depending on the application) versus error rate for the (top) reidentification and (bottom) tracking scenarios using the CRP dataset. Our method (IDBoost) operates just as fast as L2 and KISSME but has the best performance for the reidentification scenario. In the tracking scenario our method still achieves real-time operation with the best performance. This could be faster if a motion model was included. MeanShift is exceptionally fast for this reason but it's performance is poor. All experiments were conducted on a single core of a 3.20 GHz processor. The ideal performance is in the bottom left corner of the plot.

individual detectors are now being applied to every window rather than just the windows that the category detector fires on. However, this extra cost is still very small since our method utilises the features that have already been computed by the category detector.

Experiments were carried out on both the FPOQ and the CRP datasets to test this mode of operation. In this experiment sequences of individuals (a consecutive set of frames in which an individual occurs) were extracted and the category detector is evaluated on the first frame of the sequence. The output of the category detector is used to initialise an individual detector created using our method. The individual detector is then evaluated on the remaining frames in the sequence. This is a form of tracking by repeated detection using an appearance model. A motion model is not incorporated (it would be easy to implement this and would further reduce the additional runtime cost, but it would risk confusing the results of the experiments) and so the individual detector is evaluated on every window in a frame. Performance is measured by the number of times the tracker misses the individual it has been trained to track.

We compare the performance of our tracker to two other tracking methods: the Semi-Supervised Online Boosting Tracker [25] (OnlineBoost) (code obtained from authors website); and the Mean Shift or Kernel-based object tracker [9] (using the implementation in Dollar's toolbox). Both methods are initialised with the category detector output. Our method only uses the first instance of an individual from the first frame of a sequence whereas the other methods update the model of the individual over time.

The results in Figure 8 indicate that our method (IDBoost) has the best performance in terms of miss rate. Performance could be further improved by allowing the IDBoost tracker to update based on the appearance of the individual at the current frame rather than just using the appearance of the individual in the first frame of the sequence. Figure 9 (bottom) also shows that the additional computational cost of IDBoost is reasonable since real-time operation is still possible even without a motion model. Videos of our results can be found in the supplementary materials.

5. Discussion and Conclusions

We presented a method for training detectors of individual objects from a boosted category detector. Training happens in real-time using a single instance of an individual as a positive training example. The individual detectors make use of the category detector's feature computations; the thresholds for a single weak classifier are set using transfer learning. This ensures that the additional training and run-time costs for the individual detectors are minimal.

We carried out experiments on four datasets containing faces and pedestrians. The experiments were designed to test whether our simple and inexpensive strategy would work on real-world videos. We carried out two experiments: the first, designed to test reidentification, where the same individual is discovered across an entire video or image collection. The second, designed to test tracking, where an individual is tracked across consecutive video frames.

Our experiments suggest three conclusions: (a) both training and runtime computation of individual detectors is extremely inexpensive; (b) our method has both better tracking and reidentification performance than previ-

ous methods on the FPOQ and CRP datasets; (c) the cross-dataset performance of our method is better than KISSME [32], a state-of-the-art, reidentification method.

Since our results show that individual object detectors can be trained quickly it suggests that a tracking system robust to drift, could be implemented. In this system, individual object detectors are used to track individuals and are updated using the appearance of the individual on a frame-by-frame basis rather than only using the first example of the individual, as is done in this work.

Acknowledgments

This work is funded by the ARO/JPL-NASA Stennis grant NAS7.03001 and the ONR MURI Grant N00014-10-1-0933.

References

[1] R. Appel, T. Fuchs, P. Dollár, and P. Perona. Quickly boosting decision trees-pruning underachieving features early. In *ICML*, 2013. 3

[2] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool. Pedestrian detection at 100 frames per second. In *CVPR*, 2012. 2

[3] J. Berclaz, E. Turetken, F. Fleuret, and P. Fua. Multiple object tracking using k-shortest paths optimization. *PAMI*, 33(9):1806–1819, 2011. 1

[4] L. Bourdev and J. Brandt. Robust object detection via soft cascade. In *CVPR*, 2005. 2

[5] X. Burgos-Artizzu, D. Hall, P. Perona, and P. Dollár. Merging pose estimates across space and time. In *BMVC*, 2013. 1

[6] M. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *ECCV*, 1998. 1, 2

[7] D. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, 2011. 1, 2

[8] B. Coifman. Vehicle re-identification and travel time measurement in real-time on freeways using existing loop detector infrastructure. *Transportation Research Record: Journal of the Transportation Research Board*, 1643(1):181–191, 1998. 1

[9] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *PAMI*, 25(5):564–577, 2003. 2, 7

[10] M. Cristani, S. Gong, and S. Yang, editors. *First International Workshop on Re-Identification*, October 2012. 1, 2

[11] J. Davis, B. Kulis, S. Sra, and I. Dhillon. Information-theoretic metric learning. In *ICML*, 2007.

[12] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian recognition with a learned metric. In *ACCV*, 2010.

[13] P. Dollar, S. Belongie, and P. Perona. The Fastest Pedestrian Detector in the West. In *BMVC*, 2010. 2, 5

[14] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, 2009. 2

[15] A. Ess, B. Leibe, and L. Van Gool. Depth and appearance for mobile scene analysis. In *ICCV*, 2007. 5

[16] M. Everingham and et al. The 2005 pascal visual object classes challenge. In *MLCW*, pages 117–176, 2005. 1

[17] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is buffy – automatic naming of characters in tv video. In *BMVC*, 2006. 1

[18] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani. Person re-identification by symmetry-driven accumulation of local features. In *CVPR*, 2010. 6

[19] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010. 1, 2

[20] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003. 2

[21] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 22:67–92, 1973. 2

[22] Y. Freund and R. R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997. 2, 6

[23] H. Grabner and H. Bischof. On-line boosting and vision. In *CVPR*, 2006. 2

[24] H. Grabner, M. Grabner, and H. Bischof. Real-time tracking via on-line boosting. In *BMVC*, 2006. 2, 6

[25] H. Grabner, C. Leistner, and H. Bischof. Semi-supervised on-line boosting for robust tracking. In *ECCV*, 2008. 2, 6, 7

[26] D. Gray and H. Tao. Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In *ECCV*, 2008. 5

[27] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *ICCV*, 2009.

[28] O. Hamdoun, F. Moutarde, B. Stanciulescu, and B. Steux. Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences. In *ICDSC*, 2008. 1, 2

[29] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person Re-identification by Descriptive and Discriminative Classification. In *SCIA*, 2011.

[30] M. Hirzer, P. Roth, and H. Bischof. Person re-identification by efficient impostor-based metric learning. In *AVSS*, 2012.

[31] Z. Kalal, K. Mikolajczyk, and J. Matas. Tracking-learning-detection. *PAMI*, 34(7):1409–1422, 2012. 2

[32] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof. Large scale metric learning from equivalence constraints. In *CVPR*, 2012. 6, 8

[33] M. Lades, J. Vorbruggen, J. Buhmann, J. Lange, C. von der Malsburg, R. Wurtz, and W. Konen. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993. 2

[34] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[35] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1, 2

[36] A. Mignon and F. Jurie. PCCA: A new approach for distance learning from sparse pairwise constraints. In *CVPR*, 2012.

[37] B. Prosser, W.-S. Zheng, S. Gong, and T. Xiang. Person Re-Identification by Support Vector Ranking. In *BMVC*, 2010.

[38] D. Ramanan, D. A. Forsyth, A. Zisserman, and S. Member. Tracking People by Learning Their Appearance. *PAMI*, 29(1):65–81, 2007. 1

[39] M. Riesenhuber, T. Poggio, et al. Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2:1019–1025, 1999. 2

[40] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof. Mahalanobis Distance Learning for Person Re-identification. In *Person Re-Identification*, chapter 12, pages 247–267. 2014. 6

[41] W. Schwartz and L. Davis. Learning discriminative appearance-based models using partial least squares. In *CGIP*, 2009. 5

[42] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001. 2

[43] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004. 2

[44] K. Q. Weinberger and L. K. Saul. Fast solvers and efficient implementations for distance metric learning. In *ICML*, 2008.

[45] L. Wolf, T. Hassner, and Y. Taigman. The one-shot similarity kernel. In *ICCV*, 2009. 6

[46] W.-S. Zheng, S. Gong, and T. Xiang. Person re-identification by probabilistic relative distance comparison. In *CVPR*, 2011. 1, 2